Problemas de Classificação

TUANY DE PAULA CASTRO

Problema: classificação

- Problemas similares ao de predição em regressão;
- **Objetivo**: com um amostra de observações independentes, construir uma função **g** que possa predizer novas observações;
- **Diferença**: variável resposta **Y** qualitativa;
- **Exemplo**: prever se um paciente tem certa doença

com base em suas variáveis clínicas (**x**).

Avaliação do método de classificação

- Como avaliar a assertividade de um método de classificação?
- O risco de um método é a probabilidade de erro em uma nova observação:

$$\mathbb{P}[Y \neq g(x)]$$

- **Resultado**: a melhor função de classificação é aquela que classifica Y na classe onde há maior probabilidade *a posteriori;*
- Tal classificador é conhecido como classificador de Bayes.

Avaliação do método de classificação

- O risco do método pode ser estimado utilizando *data splitting* ou validação cruzada;
- ***Data splitting***:
  - **Passo 1:** Usamos o conjunto de treinamento para estimar a função de classificação;
  - **Passo 2:** Usamos o conjunto de validação para calcular o risco (proporção de erros).
- **Seleção de um método de classificação:** podemos escolher pelo modelo com menor risco estimado.

Avaliação do método de classificação

- Nem sempre o risco traz toda a informação sobre quão razoável a função de classificação **g** é;
- Por exemplo, suponha que **Y** indica se uma pessoa tem certa doença rara e que, portanto, numa amostra independente, há poucos pacientes com **Y = 1.** O indicador **g(x) = 0** terá risco baixo, pois $\mathbb{P}[Y \neq 0]$ é pequena, mas sua performance deixa a desejar (classificará a maioria como zero);
- **Matriz de confusão:**

	Valor verdadeiro	
Valor Predito	Y=0	Y=1
Y=0	VN	FN
Y=1	FP	VP

Avaliação do método de classificação

- **Sensibilidade**: dos pacientes doentes, quantos foram corretamente identificados? Fórmula: S = VP/(VP + FN)
- **Especificidade**: dos pacientes não doentes, quantos foram corretamente identificados? Fórmula: E = VN/(VN + FP)
- **Valor preditivo positivo**: dos pacientes classificados como doentes, quantos foram corretamente identificados? Fórmula: VPP = VP/(VP + FP)
- **Valor preditivo negativo**: dos pacientes classificados como não doentes, quantos foram corretamente identificados? Fórmula: VPN = VN/(VN + FN)

Avaliação do método de classificação

- No caso da doença rara, a sensibilidade será próxima de zero e a especificidade será próxima de 1;
- É importante olhar outras medidas além do risco, principalmente em estudos desbalanceados;
- A sensibilidade estima $\mathbb{P}[g(x) = 1 \,|Y = 1]$;
- A especificidade estima $\mathbb{P}[g(x) = 0 \mid Y = 0]$.

Classificadores *Plug-in*

- Método simples para resolver o problema de classificação:
  - Estimar $\mathbb{P}[Y = c \,|x]$ para toda possível classe c;
  - Escolher a classe cuja probabilidade *a posteriori* é maior.

- Esta abordagem é conhecida como classificador *plug-in,* pois pluga-se o estimador da probabilidade condicional na fórmula do g ótimo;

- Como estimar $\mathbb{P}[Y = c \,|x]$*?*

Métodos de regressão

- **Regressão Linear**

Para cada classe c, estimar:

$$\mathbb{P}[Y = c\,|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

- Podemos estimar os coeficientes do modelo pelo Método de Mínimos Quadrados;
- A classe escolhida será aquela que maximiza a probabilidade;
- Problema: estimativas podem ser menores do que zero ou maiores do que um.

Métodos de regressão

- **Regressão Logística**

- Y deve assumir apenas duas possíveis classes (binário):

$$Y|x \sim Bernoulli(p)$$

$$p = \mathbb{P}[Y = 1\,|x] = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$

- Estimação pelo Método de Máxima verossimilhança por meio de algoritmos numéricos;

- No R, podemos utilizar a função *glm*.

Métodos de regressão

- **Análise Discriminante**

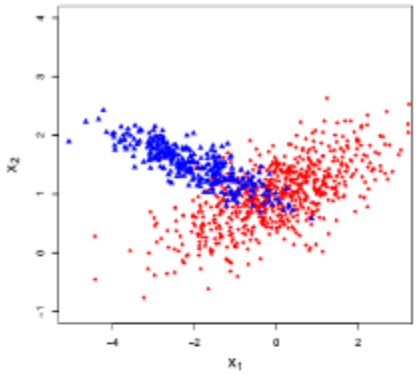
- Baseado no Teorema de Bayes:

$$\mathbb{P}[Y = c|x] = \frac{f(x|Y = c)\mathbb{P}[Y = c]}{f(x)}$$

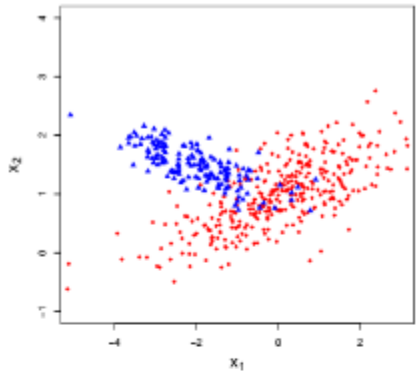
- Supõe-se que $X|Y = c \sim Normal$ e assim as distribuição $f(x|Y = c)$ e $f(x)$ são conhecidas;

- Existem duas formas:
  - **Análise Discriminante Linear:** Distribuições Normal com mesma matriz de variâncias-covariâncias (médias podem ser distintas);
  - **Análise Discriminante Quadrática:** Distribuições Normal com médias e matrizes de variâncias-covariâncias distintas.

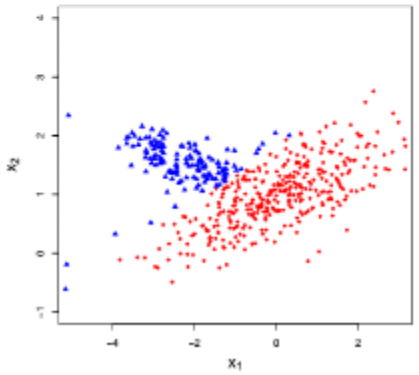
Métodos de regressão



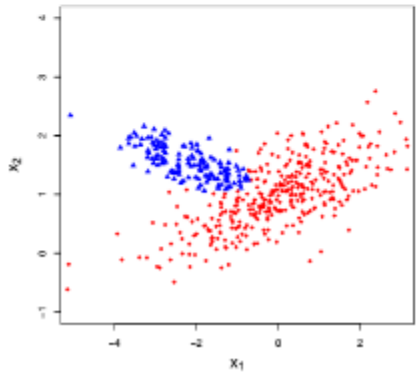
(a) Amostra de Treinamento



(b) Amostra de Teste



(c) Análise Discriminante Linear



(d) Análise Discriminante Quadrática

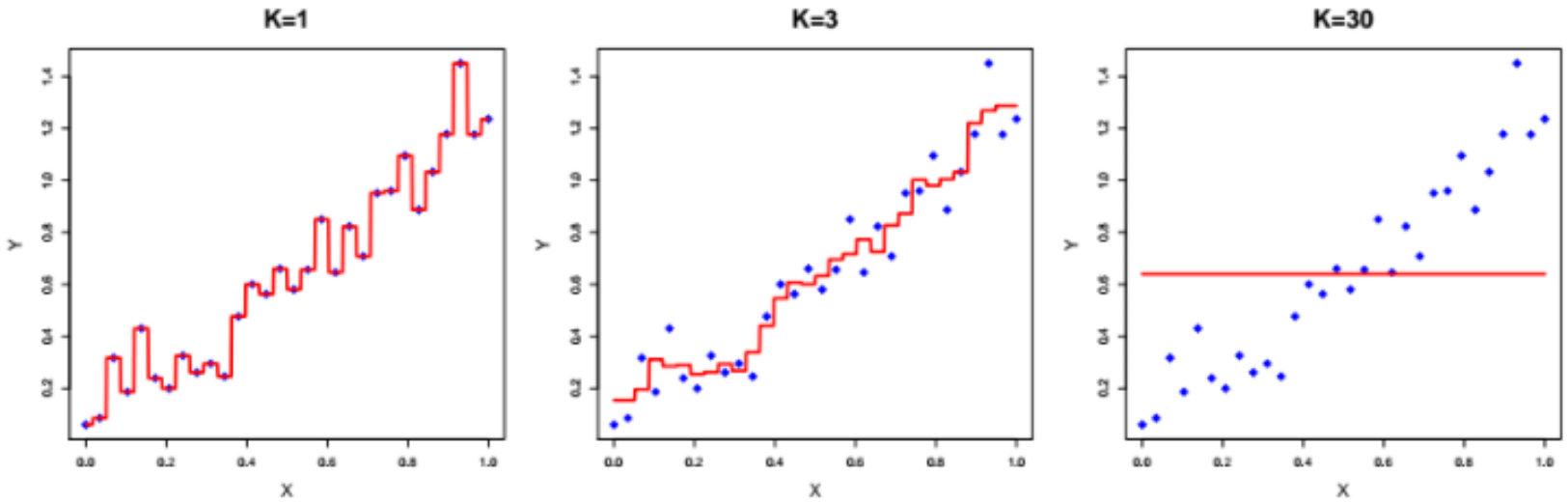
Árvores de Classificação

- Método não paramétrico com base em particionamentos recursivos no espaço das variáveis explicativas;
- Cada particionamento é um nó e cada resultado é uma folha;
- **Utilização**: verificamos se a primeira condição (nó no topo) é satisfeita; caso seja, seguimos à esquerda, caso contrário, seguimos à direita. Prosseguimos até atingir uma folha;
- **Passo 1 – Criação:** as partições são criadas buscando minimizar a proporção de erros em cada folha;
- **Passo 2 – Poda:** Cada nó é retirado por vez e observa-se a proporção de erros de predição na amostra de validação. Decide-se então quais nós permanecerão na árvore;
- Ramos grandes indicam uma importância maior da explicativa na predição da resposta.

Método dos k vizinhos mais próximos

- Um dos mais populares na comunidade de aprendizado de máquina, chamado KNN;
- Associa-se à nova observação a classe mais frequente dentre os seus k vizinhos mais próximos (distância euclidiana);
- O parâmetro k pode ser escolhido por validação cruzada;
- Um valor alto de k leva a um modelo muito simples, com uma variância baixa porém um viés muito alto;
- Um valor baixo de k, por sua vez, leva a um modelo com viés baixo, porém variância alta.

Método dos k vizinhos mais próximos



Influência na escolha de k no estimador dos k vizinhos mais próximos.

Exercícios

1) Em *dados_cobras.xlsx,* temos medidas referentes ao gênero, idade, comprimento da cauda (em mm) e o comprimento do focinho (em mm) de cobras aquáticas.

(A) Construa o gráfico de dispersão com o comprimento da cauda no eixo horizontal e o comprimento do focinho no eixo vertical e use diferentes símbolos para indicar os gêneros feminino e masculino. Parece que essas duas variáveis discriminam os gêneros?

(B) Assumindo matrizes de covariâncias iguais, encontre as funções discriminantes de Fisher e classifique uma cobra com 3 anos de idade, 140 mm de cauda e 500 mm de focinho.

(C) Assumindo distribuição Normal, faça a análise discriminante e classifique a cobra do item anterior.

(D) Qual dos dois métodos tem maior proporção de acertos?

Exercícios

2) Em *vinhos.csv* encontram-se os resultados de uma análise química de vinhos produzidos na mesma região da Itália, mas derivados de três cultivadores diferentes. A análise determinou as seguintes 13 variáveis para cada um dos três tipos de vinhos: Álcool, Ácido Málico, Cinzas, Alcalinidade das cinzas, Magnésio, Fenóis totais, Flavonoides, Fenóis não flavonoides, Proantocianinas, Intensidade da cor, Matiz, OD280/OD315 e Prolina.

- (a) Qual variável parece diferenciar mais os cultivadores de vinhos?
- (b) Aplicando uma Análise Discriminante, quantas funções discriminantes podem ser obtidas com esses dados?
- (c) Quais variáveis explicativas são significativas para a discriminação dos grupos? Explique.
- (d) Qual o percentual da variância que é explicada pelas funções discriminantes?

Exercícios

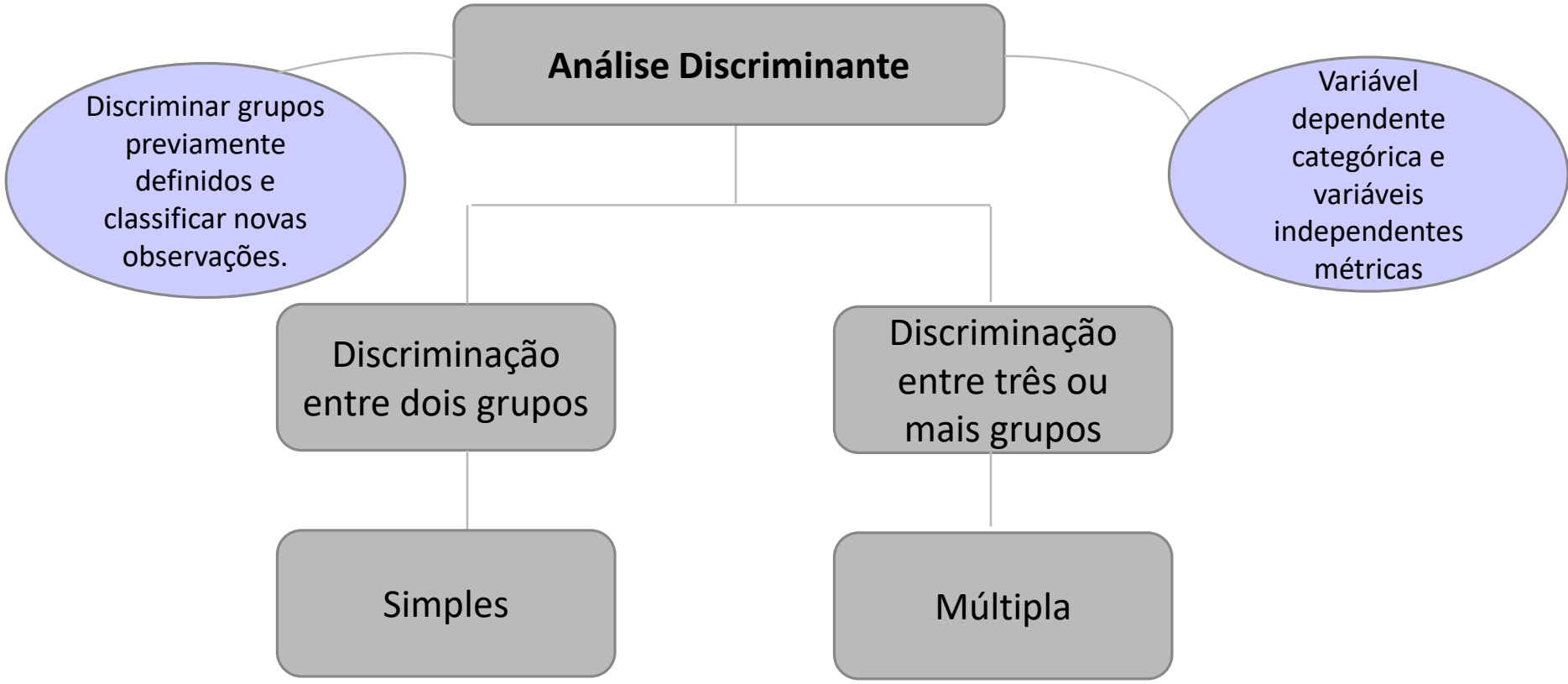
- (e) Avalie e comente a significâncias das funções discriminantes.
- (f) Construa e analise o gráfico scatterplot com as funções discriminantes.

3) No arquivo *dados_doenca.xls*, encontram-se dados provenientes de um estudo cuja finalidade é identificar fatores de risco para a doença coronariana (definida como obstrução de mais de 50% de pelo menos uma coronária). Compare os modelos de classificação estudados para prever se um paciente tem ou não doença (LO3) a partir das variáveis explicativas: presença de angina estável (ANGEST), antecedentes hereditários (AH), infarto do miocárdio prévio (IMP), nível de triglicérides para pacientes sem medicamento (TRIGS), nível de colesterol para pacientes sem medicamento (COLS), idade (IDADE1) e sexo (SEXO). Considere somente os participantes com dados completos para as variáveis indicadas.

Ajudas

- Quick R:
  http://www.statmethods.net/index.html
- Statistics with R:
  http://zoonek2.free.fr/UNIX/48_R/all.html
- R tutorials (William B. King):
  https://ww2.coastal.edu/kingw/statistics/R-tutorials/index.html
- R bloggers:
  https://www.r-bloggers.com/
- R bloggers (Brasil):
  https://github.com/marcosvital/blogs-de-R-no-Brasil

Resumo


Análise Discriminante
Discriminar grupos previamente definidos e classificar novas observações.
Variável dependente categórica e variáveis independentes métricas
Discriminação entre dois grupos
Discriminação entre três ou mais grupos
Simples
Múltipla

Referências

- BERENSON, Mark L; STEPHAN, David; LEVINE, David. Estatística: teoria e aplicações usando Microsoft excel em português. 3 ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos, 2005.
- BISQUERRA, R; CASTELLA, J.; VILLEGAS, F. Introdução à estatística: enfoque informático com o pacote estatístico SPSS. Porto Alegre: Artmed, 2007.
- FAVERO, L.P.; BELFIORE, P.; SILVA, F.; CHAN, B. Análise de Dados: Modelagem Multivariada para Tomada de Decisões. Rio de Janeiro, 2010, Editora Campus.
- HAIR, J.; ANDERSON, R.; BLACK, W. Análise multivariada de dados. 5 ed. Reimp. Porto Alegre: Bookman, 2007.
- JOHNSON, R. and WICHERN, D. Applied Multivariate Statistical Analysis. Sixth edition, Wisconsin, Pearson, 2007.